Australian Government
Department of Defence
Defence Science and
Technology Organisation

DSTO

Using Speech Technology to Improve Transcriptions: An Exploratory Study

Alex Yates, Ashley Cook and
Ahmad Hashemi-Sakhtsari

DSTO-GD-0399

Australian Government
Department of Defence
Defence Science and
Technology Organisation

Using Speech Technology to Improve Transcriptions: An Exploratory Study

Alex Yates, Ashley Cook and Ahmad Hashemi-Sakhtsari

Command and Control Division
Information Sciences Laboratory

DSTO-GD-0399

ABSTRACT

Modern speech technology is finding many new application areas within Defence. Transcription is one area where speech recognition technology is starting to replace manual methods. This adoption of speech technology is motivated by its potential to save transcribers significant amounts of time and physical effort. As part of their evaluation of military organisations, analysts from the Theatre Operational Analysis (TOA) Group within the Command and Control Division (C2D) transcribe recorded information that has been captured during interviews with Defence personnel. An exploratory study was conducted within TOA Group to look at ways of using commercial speech technology to assist in the transcribing of recorded interview material. This report describes the method used and the results obtained from that study. The report also compares traditional manual transcription approaches with newer approaches that make use of speech recognition technology. The qualitative and quantitative results obtained from this study will help to benchmark the utility of current commercial speech technology used for transcription.

RELEASE LIMITATION

Approved for Public Release

Published by

DSTO Information Sciences Laboratory
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: (08) 8259 5555
Fax: (08) 8259 6567

AR-013-068
March 2004

APPROVED FOR PUBLIC RELEASE

Using Speech Technology to Improve Transcriptions: An Exploratory Study

Executive Summary

Modern speech technology is finding many new application areas within Defence and transcription is one area where speech recognition technology is starting to replace manual methods. This adoption of speech technology is motivated by its potential to save transcribers significant amounts of time and physical effort. For the purposes of this report, transcription is the process of converting speech, usually captured using an audio tape recorder, into text.

As part of their evaluation of military organisations, operational analysts from the Theatre Operational Analysis (TOA) Group within the Command and Control Division transcribe recorded information that has been captured during interviews with Defence personnel. To help improve the efficiency of their transcription processes an explorative study was conducted within TOA Group to look at ways of using commercial speech technology to assist in the transcribing of recorded interview material. This report describes the method used and the indicative results obtained from the study.

Before looking at ways of using speech technology to improve existing transcription processes, it was necessary to measure and benchmark the quality of the existing manual method of transcription being used by TOA Group. Two quality measures were used: speed and accuracy. Speed was determined by measuring the time taken to transcribe a given amount of speech, and the accuracy was determined by counting the number of words correctly transcribed. Manual checking and automatic measurement tools were used to assess the transcribed text for errors, which were of three basic types: substitutions, deletions, and insertions. To assist the comparison of manual and automatic methods that make use of speech technology, a reference audio signal and a matching reference text was produced.

Once a benchmark had been established for the manual transcription method, transcription with the assistance of a state of the art commercial speech recogniser (Dragon NaturallySpeaking (DNS)) was investigated. For this part of the study the input devices, keyboard and mouse, were replaced with a microphone. The results were generally comparable with the manual approach, but were largely dependent on the typing abilities of the operators. However, the DNS speech recogniser system did significantly reduce the amount of typing that was required.

To achieve better results it was necessary to train DNS to recognise the unique voice characteristics of each operator. Training took about 12 minutes and consisted of the operator speaking into a microphone the words written in a prepared script. The training helps the speech recogniser adapt to the particular voice nuances of the

speaker. Errors in matching the voice with the text can be corrected immediately they are flagged. This training is usually a straightforward process when an operator is used, but it poses some difficulties for recorded material, as there is no opportunity to correct errors.

There are many factors that can influence the number of words recognised by speech recognisers. In the study, only a few of these were examined in some detail:

- length of the pause between phrases and sentences;
- rate of speaking;
- signal to noise ratio (SNR);
- impact of training;
- computer speed;
- computer memory size; and
- dictionary type.

The study examined the intrinsic (without training) ability of DNS to correctly recognise words and investigated the influence that the different factors (above) have on the performance of DNS. Training using recorded material was conducted and some of the factors were combined so as to increase the number of words correctly recognised. Combining the factors produced a significant improvement in the speech recognition results (just over 92% of the words were correctly recognised) when testing using the material on which DNS had been trained. However, for testing using new (unseen) material there was no improvement in the number of words recognised over the intrinsic results (which were around 60%).

A high SNR for the recorded material was also seen as important in increasing the number of correct recognitions. A high SNR enables DNS to keep internal search times to a minimum, which allows it to find words quickly and to keep up with the incoming audio.

By intent, this study was exploratory in nature and lacked the rigour of a formal scientific experiment. The sample sizes were too small to give conclusive results, however, they were sufficient to give indicative results. Future studies, that aim to establish the effectiveness of speech technology in assisting transcriptions, would benefit from repeating some of the investigations carried out in this study, but using larger sample sizes. Larger sample sizes would help to validate the results.

This study was designed to indicate the extent to which speech recognition technology can assist TOA Group in the transcription of data collected during the evaluations of military organisations. Although the results are only indicative, they show that with well-trained operators there may be some advantage to be gained by adopting speech technology. However, the technology is not yet mature enough to support the fully automatic transcriptions of recorded material without significant human intervention. It is likely that future versions of commercial speech technology products will overcome some of the shortcomings of the current products.

Authors

Alex Yates
Command and Control Division

Alex Yates graduated in Electrical Engineering from the University of Adelaide in 1981. Upon graduation, he worked at the Australian Broadcasting Corporation as a television engineer on the design and development of remote control camera systems and television studio equipment. He joined DSTO in 1987 where he managed the design and development of fire control systems and associated test equipment for the Nulka programme. He has provided systems engineering support to a number of major Defence projects and has carried out research into systems characterisation and modelling approaches for C3I. He is currently conducting research into ways of improving the effectiveness of operational analysis techniques. His other research interests include systems analysis and organisation concepts.

Ashley Cook
Command and Control Division

Ashley Cook is a Technical Officer (Engineering) at DSTO Salisbury. After completing a radio apprenticeship at DSTO (WRE) in 1966, he worked at Ranges Group, Range 'E', Woomera, in the field of timing and communications for 14 years. In 1980, he returned to DSTO Salisbury and worked for Trials Branch and Combat Systems Division, WSRL, engaged in development work associated with timing and audio recording techniques for field trials and exercises. He is currently engaged in work to facilitate the analysis of data collected from C2 organisations using speech technology.

Ahmad Hashemi-Sakhtsari
Command and Control Division

Ahmad Hashemi-Sakhtsari is a research scientist in Human Systems Integration Group. His current research work is focused on application of commercial language technology to military systems and on studying human computer interaction through speech as well as manual modalities. He leads a small speech technology research and development team in the Human Systems Integration Group.

CONTENTS

Figures

Tables

Abbreviations

ADF	Australian Defence Force
C2	Command and Control
C2AST	C2 Australian Theatre (Group)
C2D	Command and Control Division
TOA	Theatre Operational Analysis (Group)
CCIS	Command Control Information Systems (Branch)
CEP	Cool Edit Pro
CPU	Central Processing Unit
db	Decibel
dbA	Decibel Acoustic
DNS	Dragon NaturallySpeaking
DSTO	Defence Science and Technology Organisation
HIS	Human Systems Integration
MHz	Mega Hertz
MS	Microsoft
NIST	National Institute of Standards and Technology (US)
PC	Personal Computer
RAM	Random Access Memory
SCTK	Scoring Toolkit (Speech Recognition)
SNR	Signal to Noise Ratio
VU	Volume Units
w/s	Words per second

1. Introduction

This section introduces the context for the report. It provides background motivation for the work and identifies its intended purpose and scope. An outline of the contents of each of the sections of the report is also provided.

1.1 Context

Theatre Operational Analysis (TOA) Group is concerned with improving the effectiveness and efficiency of the processes it uses to evaluate military command and control (C2) organisations. The research efforts within TOA Group are focused on making better use of automatic tools to reduce the amount of effort expended in all the phases of the evaluation process. Of particular value to the analysts who work within TOA Group are those tools that can help to collect, transform, visualise, and make sense of the data obtained during interviews with personnel who work at the operational level within the various military headquarters. These include tools that can assist analysts in processing and making sense of the spoken word.

This report describes the method and results associated with a pilot study carried out within TOA Group that investigated the extent to which current speech technology could assist operational analysts in their evaluations of military organisations. From the onset, the study was explorative in nature and lacked the rigour of a formal scientific experiment. Also, because of the small sample sizes used for the study the results are not conclusive, only indicative.

The intended readers of this report are the operational and organisational analysts within the Command and Control Division. In particular, analysts within:

- Theatre Operational Analysis (TOA) Group;
- C2 Australian Theatre (C2AST) Group;
- Human Systems Integration (HSI) Group; and
- Speech researchers within Command Control Information Systems (CCIS) Branch.

The report may also find readership in other operational analysis areas within the Defence Science and Technology Organisation (DSTO).

1.2 Background

TOA Group, which is a part of the Command and Control Division (C2D), provides advice to the ADF on ways of improving the effectiveness of military headquarters' organisations. To provide this advice TOA conducts a range of evaluations that examine military organisational structures, business processes, and associated military information systems. These evaluations examine the ability of organisational

architectures to effectively support operational level military commanders, and they focus on the people, the tasks, the resources, and the relationships that make up military organisations. An important aspect of these evaluations includes assessing the impact of the introduction of newer information technologies on operational performance.

1.3 Evaluation process

Typical organisation evaluations involve five phases: planning, data collection, data reduction, analysis, and reporting.

1.3.1 Planning

During the planning phase the scope of the evaluation is determined. The problem to be investigated, the purpose of the evaluation, and the terms of reference are agreed with the client and a detailed work program is prepared. It is during the planning that the practical issues regarding the type and quantity of the organisational data that are to be collected are determined.

1.3.2 Data collection

Data are collected using various methods including observation, questionnaires, and interviews. This report focuses on aspects of the post-processing of speech data collected during interviews. Typical interviews take about 60 minutes and during the interviews, the salient parts of what is said is usually noted by the interviewer using a pen and paper, while the full conversation, if security restrictions allow, is captured using a tape recorder.

1.3.3 Data reduction

During data reduction, the interview data is converted into a form that is suitable for further analysis. If audio recordings are made, it is usually necessary to transcribe the recorded speech into text in order for it to be suitable for further analysis. A particularly onerous task is that of transcribing recorded interview material into text. Speech tools that can speed-up this transcription process have the potential to shorten the total time required to carry out the data reduction.

1.3.4 Analysis

The analysis phase aims to identify the important properties and relationships that exist within the organisation being evaluated. During this phase the information collected during the client interviews is filtered and sorted into categories. Modelling tools are employed to help manipulate and visualise the information and extract important issues. The results from the analysis are used to make recommendations.

1.3.5 Reporting

During the course of an evaluation, presentations and reports that describe the work progress, results, issues, and recommendations, are given to the clients.

1.4 Purpose of this report

The purpose of this report is to indicate, in a practical way, to what extent speech technology can be used to improve the efficiency of the transcription process used during the evaluations of military organisations. This report describes a pilot study that was conducted by TOA Group that compares manual transcription processes with automatic transcription processes. The study identified some of the salient factors that can be adjusted in order to improve recognition performance using speech technology, and the report makes recommendation regarding the direction of future studies.

1.5 Scope

The report identifies and discusses some important issues associated with the use of speech technology and how it can be used to improve the evaluations of military organisations. However, the study is exploratory in nature, and its findings, although of immediate use to TOA Group, are really intended to scope future studies. The experimental rigour has been relaxed to allow a wider exploration of some of the issues involved. Future studies are needed to validate the results obtained from the approaches used in this study. The recommendations in section 14 give some guidance regarding the future directions of such studies.

1.6 Overview

Section 1 outlines the context, background, purpose, and scope of this report.

Section 2 discusses in general terms current speech recognition technology.

Section 3 gives an outline that includes the aim and method used to conduct the study.

Section 4 describes how a basic reference text was selected and prepared, and how the audio recording and the matching reference text were derived from that basic reference text.

Section 5 describes how speed and accuracy metrics are used to determine speech recognition quality. The different scoring methods used during the study are also briefly described.

Section 6 examines traditional manual methods that are used for transcription. The results from this part of the study are used as a basis for comparison with the automatic methods used later in the study.

Section 7 describes and gives the results of using a computer software assisted semi-automatic transcription process to transcribe recorded speech.

Section 8 describes the process used to transcribe recorded speech using speech recognition software. A range of factors that influence the number of words correctly recognised are identified and discussed.

Section 9 looks at ways of improving the recognition of recorded speech and systematically describes modifications to the study conditions.

Section 10 describes and gives the results of combining the improvements identified in the previous section.

Section 11 examines some other ways in which the manual transcription of recorded speech can be aided by automatic transcription processes.

Section 12 examines how extending the length of the reference text impacts on speech recognition performance.

Section 13 concludes the report by summarising the salient points to come out of the study.

Section 14 makes recommendations on how to use speech recognition technology and identifies some future areas for experimentation.

2. Speech Recognition Technology

This section gives a brief overview of speech recognition technology. It describes the basic principle of operation of the speech recogniser used in this study and it identifies some of the salient factors that are known to influence speech recogniser performance.

2.1 Principle of operation

Computer based systems are now available that will allow computers to effectively recognise and respond to human speech. These systems are able to convert the continuous audio signals derived from speech into representations of basic speech patterns (phonemes and words). The process model (Talintyre 1996) for the speech recognition system used in this study is represented in Figure 2-1.

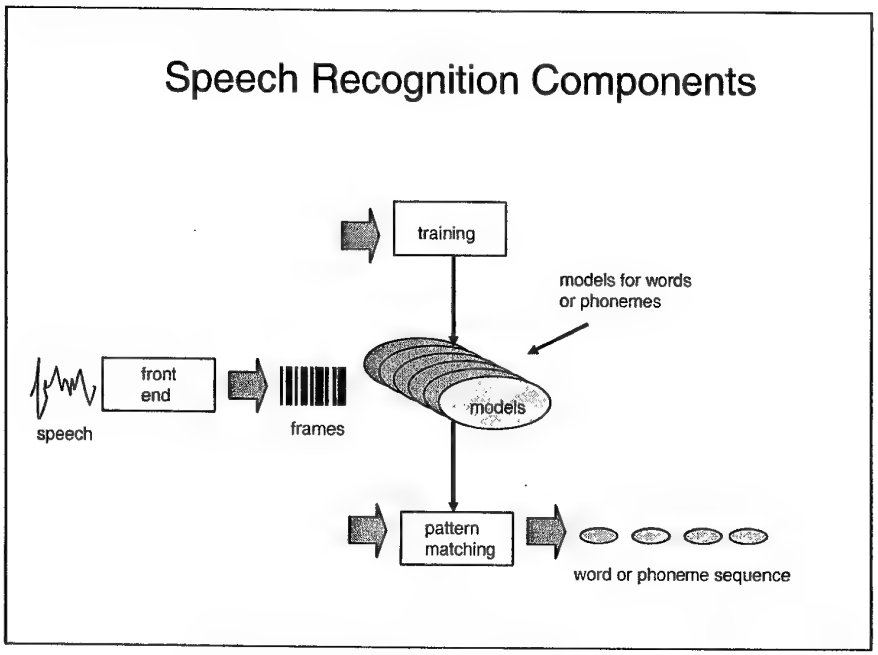

Figure 2-1 Common components of a speech recognition system

Incoming speech is sampled at regular intervals by the front-end processor and converted into frames, which represent speech in a compact and consistent way. The frames are used to build phoneme sequences (words), which are then pattern-matched against models of stored phoneme sequences. If an input pattern is recognised as being the same as a stored phoneme sequence, it generates an appropriate output from the speech recognition system in the form of a word. If the input pattern is not recognised, a search is made for the closest phoneme sequence match, which then becomes the output word. The output of this type of system is usually represented on the computer screen as text.

The models for the phonemes, which are often referred to as reference patterns, are produced during training. The training can be used to introduce new reference patterns, or to help the speech recognition system to adjust to the voice characteristics (eg accent) of a particular speaker. Many speech recognition systems come with a library of pre-programmed reference patterns (representing word sounds). To introduce new words it usually necessary to enter them via a keyboard. The new words then become part of the library. To adjust to the voice characteristics of a speaker, it is usually necessary for the user to spend a few minutes reading aloud a pre-programmed training script.

Traditionally, each reference pattern is trained on many examples of a particular word, which is transformed into a distillation of the many ways that particular word can sound. This type of training can be slow. A faster and less expensive alternative for

building reference patterns is to string phonemes together. As there are only about 50 phonemes for the English language any new reference pattern can be rapidly constructed.

2.2 Factors influencing the performance of speech recognisers

Over 80 factors have been identified as affecting the performance of speech recognisers (Lea 1982; Lea 1983; Pallett 1985). Some of the more important factors are grouped below:

2.2.1 Task related factors

Speech recognition can depend on task related factors such as:

- type of device and the interface to the rest of the system;
- imposition of syntactical constraints; and
- removal of a rejection threshold, which enforces word selection.

2.2.2 Human factors

There are many human factors issues that can influence speech recognition. These include characteristics of the speaker such as:

- gender;
- age;
- speaking rate;
- speech level;
- education and training;
- dialect history and pronunciation habits;
- particular speech idiosyncrasies;
- variability of word articulation introduced by the speaker;
- speaker generated noises such as coughs and tongue clicks;
- motivation;
- fatigue; and
- form of the speech (eg isolated, connected, or continuous).

2.2.3 Language factors

Language factors that can influence speech recognition include:

- active sub-vocabulary of the allowable next words;
- length of words in milliseconds, and the number of syllables;
- language spoken; and
- consonants and vowel patterns in speech.

2.2.4 Ambient or environmental factors

Ambient or environmental factors include:

- speech SNR;
- spectral content of the noise;
- nature of the noise;
- transmission SNR;
- transmission channel bandwidth and phase distortion;
- recording distortion; and
- type of microphone used.

2.2.5 Algorithmic factors

Algorithmic factors include:

- alignment of discrete sampling with the waveform;
- positioning of word boundary locations;
- time and amplitude normalisation procedures;
- pattern matching control strategy; and
- degree of focus on prosodic variables and speech distinguishing features.

2.2.6 Performance and response factors

Performance and response factors include:

- types of error;
- verification of a decision by auditory means or visual display;
- feedback of intermediate results; and
- procedures for correcting errors.

2.3 Dragon NaturallySpeaking (DNS)

There are several commercial brands of speech recognition system available. The choice of Dragon NaturallySpeaking (DNS) for this study over the other brands was based on the extensive use that was being made of DNS by other researchers within C2D, and the results of a survey (Vozzo 2001) of speech recognition products that was conducted as part of the Australian Defence Force Academy's Project Vocoder (Lapworth and Jager 2000).

Janet Baker started 'Project Dragon' at the Carnegie Mellon University in 1974. This research project formed the basis from which the company Dragon Systems was established in 1982. DNS is a software product that allows users to talk to their computer and have their words transcribed into documents. The study described in this report used Dragon NaturallySpeaking Professional version 4.0. However, during the time of preparing this report, versions 5.0 and 6.0 became available.

DNS is a commercial speaker-dependent large vocabulary continuous speech recogniser. Continuous speech recognisers can accept words spoken fluently, and as rapidly as in conversational speech. DNS provides dictation into any Windows application. It provides a backup dictionary containing speaker-independent spelling, along with acoustic and language information for a total vocabulary of over 270,000 words and names. An effective active vocabulary of over 160,000 words exists out of the total vocabulary. The active vocabulary is a subset of the total vocabulary that may be active in computer memory at a given time due to an imposed task grammar or other syntactic constraint. The active vocabulary can be entirely customised in that any or all of the words in its vocabulary can be replaced with other specific words.

To add words, they need to be used and spelled once. DNS attempts to generate pronunciation for new words. If the word is in the backup dictionary, DNS will conduct a search and display it. DNS automatically puts any new word into the active vocabulary so that it can be immediately recognised the next time it is uttered. It remembers spelling, acoustic patterns, and language usage.

DNS analyses incoming speech using both an acoustic model and a language model. The acoustic model is based on speech samples collected from thousands of people. Mathematical algorithms are used to compare the incoming speech to an appropriate model. The result is the best acoustic match and a short list of alternatives.

As well as British and North American vocabularies, DNS includes speech models and customised vocabularies for English speakers from the Indian subcontinent, South East Asia, and Australia.

The language model is based on an analysis of how words are used in thousands of documents. The model predicts word usage to find a better match and is designed so that DNS can distinguish between words that sound alike, such as 'to', 'two', and 'too'. A statistical process that models the likelihood of a word following others, e.g. 'little red riding hood' being more likely than 'big blue riding hood' provides enhanced performance. DNS determines the most likely word based on probability and context.

DNS adapts to each user's voice and vocabulary. A personal file is created for each user that contains words and acoustic models for both the active vocabulary and backup dictionary. As a user dictates and corrects recognition mistakes, DNS personalises the acoustic models to match that speaker. The language model is continually modified to better reflect the semantic and syntactic usage of the speaker. For instance, words that are used most frequently appear at the top of word choice lists, and newly added words replace those least frequently used.

3. Study Outline

This section explains the aim of the pilot study and gives an overview of the process that was followed. A block diagram is used to illustrate the key steps in the process.

3.1 Aim of the study

The aim of this task was to assess the effectiveness of a typical state of the art speech recognition software tool in facilitating the process of transcribing audio tapes. The study undertook the following activities:

- comparison of the speed and accuracy of manual versus computer assisted (using commercial speech recognition technology) transcription processes;
- identification of issues that significantly influence the effectiveness of speech recognition technologies that are used in transcription applications; and
- development of a report that makes recommendations regarding future studies.

3.2 Study overview

The process followed during the study is outlined in Figure 3-1. The first step in the process was to prepare a reference audio recording and associated reference text in which the words spoken in the reference audio accurately matched the words printed in the reference text. Three different methods of transcribing the reference audio were then investigated: manual transcription, DNS assisted transcription, and automatic transcription. At the completion of each transcription the transcribed material was compared to the reference text.

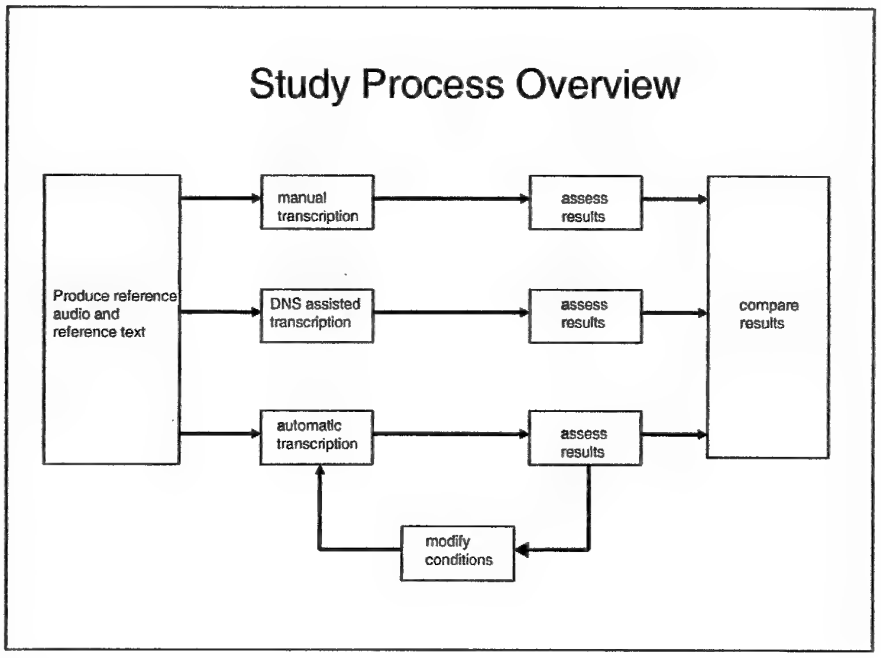

Figure 3-1 Study process overview

The first method investigated was the manual transcription method. The manual transcription method is the method that has been traditionally used within TOA Group. This method relies on the use of dictaphones and does not make use of any type of voice recognition software to assist the transcription process. The results obtained from using this method established a benchmark for comparing the effectiveness and efficiency of the other methods used in the study.

The second method investigated was the DNS assisted transcription method. The DNS assisted transcription method aligns closely with the way that the DNS speech recognition software tool is designed to be used. However, the method still makes use of a dictaphone (or another equivalent device) to allow the user to start and stop the recording. Using this approach the user listens to the recorded reference audio and then orally repeats what is heard into a microphone connected to the speech processing system, which then produces the transcribed text.

The third method investigated in this study was a fully automatic process, in which the reference audio was fed directly into the speech recognition system with a minimal amount of manual intervention. Much of this part of the study was concerned with identifying and assessing those factors that influence the number of words recognised by the speech recogniser. As there are many factors that can influence speech recognition, only a few of the more readily accessible factors were investigated, and it was necessary to modify the conditions and repeat the measurements for each factor.

Factors investigated in this study included:

- amount of training given to the speech recognition system;
- speech rate;
- SNR;
- computer speed and memory capacity; and
- type of vocabulary.

Assessment of the results for each method consisted of timing the procedures and counting the number of correct word recognitions. The transcribed text was also checked to see if the order of the words matched those in the reference text. For each method investigated the results were compared and the effects of the factors assessed.

4. Reference Audio and Reference Text

To help in the comparison of the different methods it was necessary to produce a reference audio signal and associated reference text. This section describes how the reference text and the reference audio used in this study were produced.

4.1 Purpose of the reference audio and the reference text

Three components related to the transcription process are referred to often in this report: the 'reference audio', the 'reference text', and the 'transcribed text'. The 'reference audio' is the audio input signal that is fed into the transcription system. The 'reference text' is a written text that precisely matches the reference audio. The 'transcribed text' is a written text that is produced by the transcription system (manual or automatic).

In order to assess the number of errors in a transcribed text, it is usually necessary to compare the transcribed text with corresponding reference text.

During the study, different transcription methods were compared. For these comparisons, the same reference text was used.

4.2 Producing the reference audio and the reference text

For the first part of this study a reference audio with matching reference text was produced. The basic process describing the production of the reference audio and the reference text is shown in Figure 4-1. An edited transcript of a conversation between an analyst and a military officer during a military exercise was selected as the basis for developing the reference audio. This transcript provided a representative sample of military topics, military terminology, and military acronyms. Sensitive topics and references were removed. This transcript was then modified so that it could be read as

a narrative spoken by one individual. However, the question and answer format was preserved. The transcript consisted of 2500 words and it took 12 minutes and 30 seconds to read. This corresponds to approximately 3.3 words per second, a speaking rate that is slightly higher than the average rate of 3 words per second, or 10 phonemes per second, as reported by Martin and Welch (Martin and Welch 1980). The audio recording of this transcript became the 'reference audio'. The reference audio was then carefully transcribed to produce the 'reference text'. The reference text is reproduced at Appendix A.

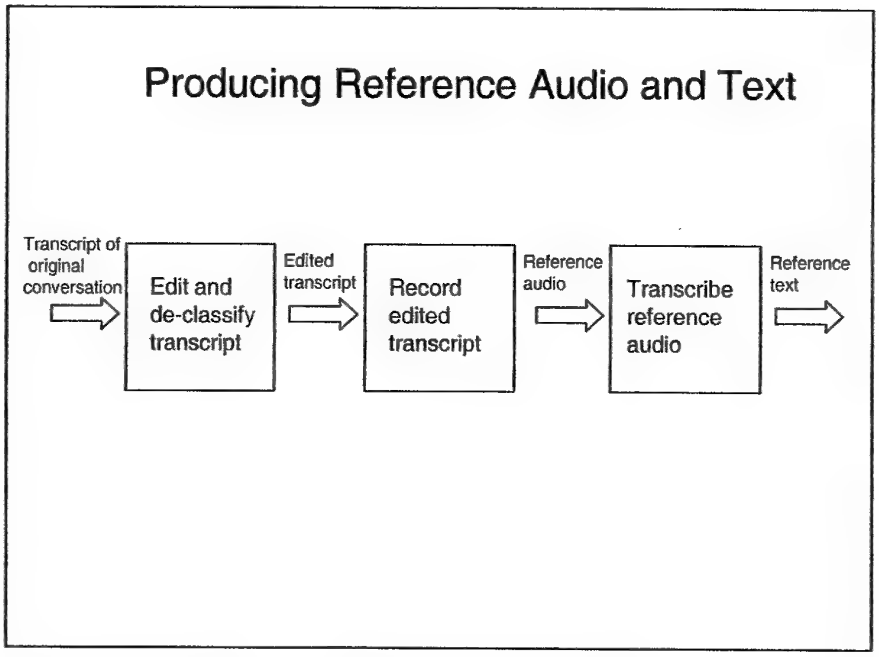

Figure 4-1 Process for producing 'reference audio' and 'reference text'

4.3 Issues in producing reference audio and reference text

4.3.1 Digital audio wave file

To provide greater flexibility in manipulating the reference audio signal, the analogue audio recorded on the tape recorder was also re-recorded as a digital audio file using Cool Edit Pro (CEP) software. CEP is a product of the Syntrillium Software Corporation.

4.3.2 Speaker selection

The speaker selected to produce the recording of the reference audio, was chosen for attributes such as, ability to read fluently, good diction, and a good understanding of

the subject matter. An Australian adult male was used to read the reference audio material. Although gender can influence speaker recognition rates, it was assumed to be constant in this study. Future studies will need to take gender issues into account. Before making the recording, no particular directions were given, except a request to read at a natural and comfortable pace as though speaking with someone.

4.3.3 Background noise level

The first recording of the reference audio took place in a conference room where the predominant background noise was that created by the air-conditioning. Four other people, who also had an interest in the study, were present during the recording, although they did not noticeably contribute to the noise level. Other noises included the intermittent sounds of doors closing, people walking along the corridors, and distant muffled speech. A background noise level of 41 dB was measured with a standard sound level meter using the 'A' weighted scale. This background noise level is typical of that for an office environment.

4.3.4 Microphone

The audio input to the recorder came from an Electret condenser omni-directional stereo microphone. The microphone, which is an accessory for the tape recorder and is typical of the type used during client interviews, was placed at a distance of approximately one metre from the speaker. The microphone placement was similar to that used during a typical client interview between two people, each seated opposite each other at a desk.

4.3.5 Recorder

The reference audio recording was produced using a 4-track, 2-channel stereo, Sony Walkman Professional tape recorder (WM-D3). The input level adjustment dial was set to the maximum (+10) in order for the input level meter to register 0VU (volume units) on audio peaks.

5. Transcription Measurements

This section describes the measures used during the study to assess the quality of transcription processes. Two types of transcription quality measures are described: speed and accuracy. The tools used to score the accuracy of the transcription are also discussed.

5.1 Transcription quality

The quality of a transcription is usually determined by measuring how long it takes a person, or machine, to produce the transcribed text from the reference audio, and by comparing how accurately the transcribed text matches the reference text. These two quality measures are usually referred to as speed and accuracy.

5.2 Speed

The speed of the transcription process is the time required to convert the reference audio into the transcribed text.

For the manual transcriptions (see section 6) the time included the time required by the participants to listen to and type the pre-recorded reference audio. As the speech rate was approximately 200 words per minute, much faster than a normal typist can type, it was necessary to stop, and in some cases repeat, the recorded material to allow each participant time to catch up. A foot operated dictaphone was used to start, stop, and rewind the tape. During the experiment, a stopwatch was used to measure the total time required to produce the transcribed text from the reference audio. Set-up times were not included.

For the automatic transcriptions the reference audio was fed directly into the speech recognition system. There was no manual intervention. The time taken for transcription did not include any set-up time. In some cases the speech recognition system was trained. However, the times measured did not include any time required for training. Various modes of automatic transcription were tried and these are described in later sections of this report.

5.3 Accuracy

The accuracy of a transcription is determined by directly comparing the reference text with the transcribed text. For comparison purposes, the differences (errors) are usually categorised and counted. Differences can be recognised between words, sentences, and paragraphs in the two texts being compared. However, most of the errors in the different texts are usually found by comparing words. The types of errors fall into one, or combinations, of the following categories:

- substitutions, where words are replaced with different words;
- deletions, where words are removed from the text; and
- insertions, where words are added to the text.

The inspection of these three categories can be used to identify errors that are caused by:

- incorrect spelling;
- incorrect punctuation;
- word insertion;

- word replacement;
- word deletion;
- word transposition;
- typographical errors;
- abbreviation errors; and
- expansion errors.

5.4 Scoring method used during the study

During the study three different methods of comparing the transcribed text with the reference text were investigated: the first made use of the MS Word 2000 text comparison feature, the second used manual scoring, and the third incorporated a software scoring program.

5.4.1 MS Word

The MS Word 2000 word processor has a text comparison feature that allows for the direct comparison of text documents. The reference text and the transcribed text can be directly compared using this feature. Limitations of this approach are discussed in the next section.

5.4.2 Manual scoring

This method involves manually comparing the reference text with the transcribed text and counting the number of errors. This is essentially a proof reading task, which can be time intensive and prone to human mistakes.

5.4.3 Scoring program

A specifically designed scoring program, Sclite, which is part of the Speech Recognition Scoring Toolkit (SCTK) version 1.2 from the US National Institute of Standards and Technology (NIST) was also used to compare the reference text and the transcribed text. The Sclite program is specifically designed for use with speech recognition systems. It compares the reference text to the transcribed text in an alignment process and is able to generate a report summarising the transcription performance. Sclite does not contain a language model for extracting meaning from the sentences. It works purely at the mechanical level of straight text comparison.

The Sclite performance report details word and sentence errors and categorises error types as substitutions, deletions, and insertions. The reporting of those errors, which enumerate the word recognition performance in terms of substitutions, deletions, insertions, and word accuracy, is generally given as a percentage of the total number of words in the reference text (Fiscus 1998). A disadvantage of this system is the need to manually insert record markers to help to group and hence synchronise the text files that are being compared. As inserting markers takes time, especially with large

documents, there is usually a trade-off between the number of markers inserted and the time taken to insert those markers. However, the smaller the word sequences between the markers, the more accurate the scoring (Kempt, Schmidt et al. 1999).

5.4.4 Transferring the NIST software from UNIX to PC

The NIST scoring software is freely available from the NIST web site on the Internet. However, as supplied it is only available for use on UNIX platforms. A public domain 'C' compiler, called DJGPP, was used by DSTO personnel (see acknowledgements) to implement changes to the UNIX based scoring software to enable it to run in a DOS environment. The validation of the NIST scoring software running in the DOS environment was sufficient for the purposes of this study, but it was not exhaustive.

6. Manual Transcription

This section examines the manual transcription method that has traditionally been used by TOA Group. The results from this part of the study form the basis for a comparison with the results obtained from the automatic transcriptions that are described later.

6.1 Manual transcription process

A block diagram of the manual transcription process is shown in Figure 6-1.

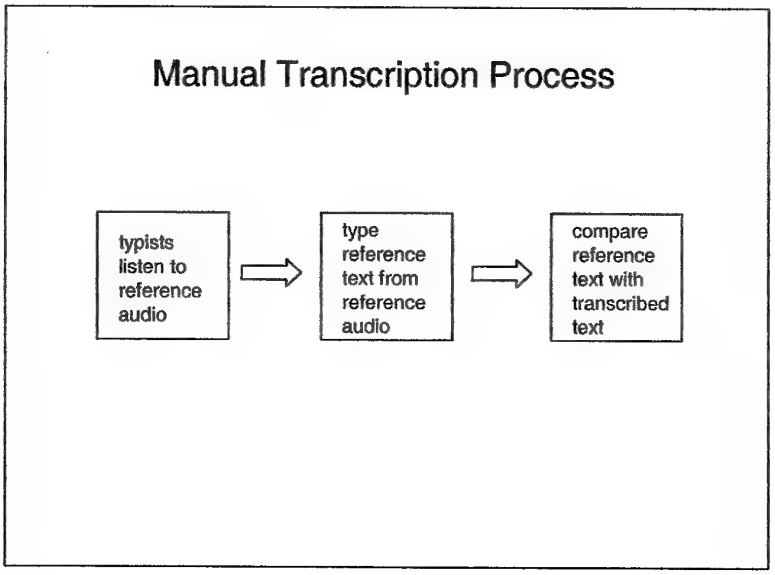

Figure 6-1 Manual transcription process

During this part of the study four participants each took turns to listen to the reference audio and, with the aid of a dictaphone, typed out the corresponding transcription text. The participants used for the study were representative of those that are actually used by TOA Group to transcribe recorded client interviews; mainly administrative support staff that work within C2D. Each of the four participants acted as an independent subject during this part of the study. They each have different amounts of expertise in keyboard skills, transcription experience, and familiarity with military terminology and acronyms. At the completion of each manual transcription, the time taken to complete the transcription was noted and the transcription text was then compared with the reference text to determine the number of words correctly recognised.

6.2 Manual transcription equipment

All of the transcripts were produced using the same computer and replay equipment. The technical details for the equipment are as follows:

- dictaphone — Sony Transcriber BM-77 (with foot control)
- headset — Yamaha Orthodynamic headphones HP-2
- computer — TPG Pentium
- software — Windows 95 version 4 and Microsoft Word 95

6.3 Manual transcription speed

Table 6-1 shows the time taken for each of the participants to transcribe the reference audio. The table also provides an indication of the keyboard skill levels, and the military terminology awareness for each of the participants. A typing skill test was not carried out, but generally, the keyboard skills of the participants were considered to be typical of that found in a modern office environment. Typing speeds were around 40-80 words per minute. Also, each participant had more than five years experience working in a military/DSTO environment, and each had a basic knowledge of military terms and acronyms.

Table 6-1 Participant speed during manual transcription

Participant	Time taken (mins)	Keyboard skills	Terminology awareness
1	70	Good	Complete
2	83	Extensive	Minimal
3	74	Good	Moderate
4	80	Good	Good

The times recorded did not include set-up times or any breaks required during the transcription process. It can be seen from table 6-1 that the time taken to produce the transcribed text ranged from approximately six to seven times the duration of the reference audio (12 minutes and 30 seconds). The time taken by the participants to complete the transcriptions may fall short of industry best practice. However, this is to be expected because the participants used in this study usually spend only a small part of their normal work time with transcribing activities.

6.4 Manual transcription accuracy

The comparison feature of MS Word was used to identify the differences between the reference text and the transcribed text. MS Word automatically underlines differences. Once MS Word had identified those differences, manual scoring was used to categorise and count the errors. Table 6-2 shows some examples of the types of errors that were made during the manual transcription process.

Table 6-2 Examples of the types of errors made during manual transcription

CATEGORY	REFERENCE TEXT	TRANSCRIBED TEXT
Abbreviation	'I have'	'I've'
Expansion	'we'll'	'we will'
Punctuation (difference)	'hour, but'	'hour. But'
Spelling	'effect mobility'	'affect mobility'
Word omission	'get all your locstats'	'get your locstats'
Character omission	'activity'	'ativity'
Insertion	'sorted out, probably adds'	'sorted out and probably adds'
Replacement	'in which'	'where'
Transposition	'throw a map out and go'	'throw out a map and go'
Typographical	'before'	'beforee'

Table 6-3 lists the number and categories of errors made during the manual transcription process. Manual scoring was used to count and categorise the errors. The total number of errors as a percentage of the total number of words in the reference text is also given. From the results it can be seen that for all of the participants used in the study the accuracy (number of correct words) is quite high (greater than 96%). As the reference audio was only 12 minutes and 30 seconds in duration, the effects of fatigue are likely to be minimal. For longer duration transcription tasks, factors such as fatigue are likely to negatively influence the number of correct recognitions.

Table 6-3 Errors for manual transcription

CATEGORY	COUNT			
	Participant 1	Participant 2	Participant 3	Participant 4
Spelling	8	0	8	3
Word omission	47	41	17	17
Character omission	8	0	4	2
Insertion	7	7	8	6
Replacement	20	9	10	5
Transposition	2	0	1	0
Typographical	18	17	10	7
Abbreviation	9	4	3	4
Expansion	2	2	3	2
Punctuation	111	115	168	89
Correct words	2399/2500	2433/2500	2451/2500	2466/2500
Correct words (%)	96%	97%	98%	98%

6.5 Comparison features of MS Word

During the manual transcription process, the participants who listened to the reference audio were unaware of the format of the reference text. This initially resulted in a very large number of differences between the reference text and the transcribed text. The differences were due to the nature of the comparison feature in MS Word, which is specifically designed to help users to compare versions of the same document and to track changes. As a result, even the smallest formatting differences between two seemingly similar documents are underlined and such things as font size, style, and punctuation differences can produce errors. To reduce these errors the reference text and the transcribed text were each saved as a text file, which removes the formatting and makes text comparison easier.

7. Transcription Using DNS

This section describes how manual transcription can be assisted through the use of DNS. This approach aligns closely with the way that the manufacturer (Dragon Systems) intended DNS to be used.

7.1 Transcription approach used with DNS

A block diagram illustrating the approach used is shown in Figure 7-1.

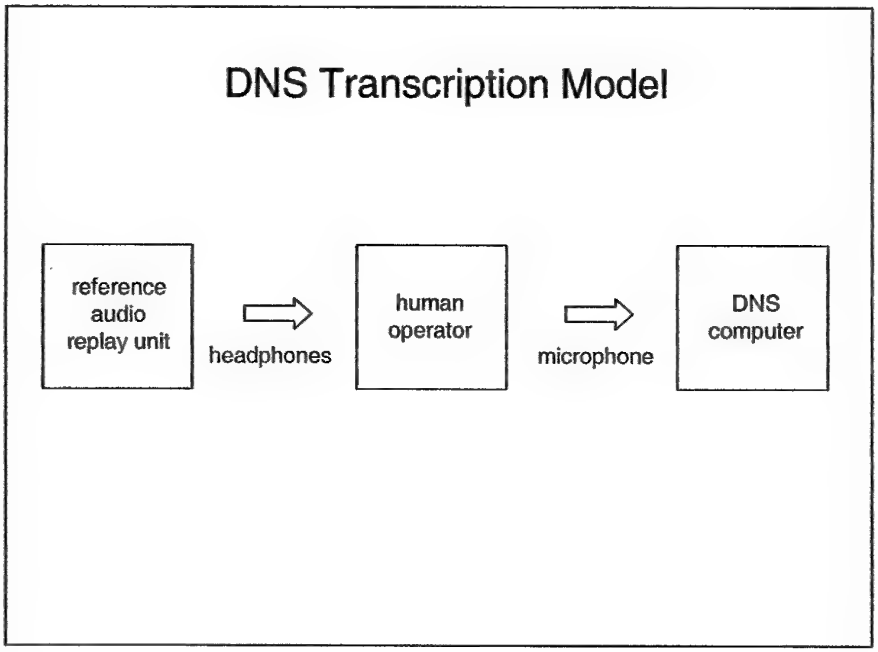

Figure 7-1 DNS transcription model

7.2 DNS training

DNS is a speech recognition software tool that is able to take voice inputs from a microphone and convert them into text, and as supplied DNS has the ability to recognise a large number of words. However, the number of words that DNS correctly recognises can be significantly increased if the user gives DNS a short period of training. This training generates speech phoneme models and adapts the language model to suit particular speakers. Another feature of DNS is that it is able to self-train while being used, which helps the recognition capability to improve with use.

Two people participated as subjects in this part of the study. For each participant, the system was trained using the 'Alice's Adventures in Wonderland' standard training script, which is supplied with DNS. During training, the text for the training script was displayed on the computer screen and read aloud by each participant into the microphone. In each case the training took approximately 12 minutes.

7.3 Transcription

During transcription, the recorded reference audio (see Appendix A) was fed into the headphones and the participants repeated what they heard into the microphone. Their speech was processed by DNS and displayed on the computer screen. The word differences between what was said and what was displayed were corrected during the transcription process. As the reference audio was recorded on tape, facilities were needed to start, stop, and rewind the tape. In this case a Sony Transcriber (Type BM-77) was used.

7.4 Results using DNS

Speed and accuracy measures, similar to those used for the manual transcription process described in the previous section, were used to assess the transcriptions carried out using DNS. Table 7-1 shows the transcription results for the two participants. Although there was a variation in the time taken to produce each transcription, the transcription accuracy remained consistently high (greater than or equal to 96%). It should be noted that during the transcription the participants were responsible for correcting any errors they observed. Again, as was done in the manual transcription case, the final accuracy checking (scoring) of the number of words correctly transcribed was independently determined by manually comparing the reference text with the transcribed text.

Table 7-1 Transcription by dictation using speech recogniser.

Participant	Time taken	Correct recognitions
No. 1	110 minutes	98%
No. 2	74 minutes	96%

Both the typing and dictation experience of the participants, and the amount of training given to the speech recogniser, are important factors to consider when conducting transcriptions in this manner. Both participants were able to type and use a word processor, but with different abilities. Neither of the participants had any previous experience with the use of DNS.

The time taken to produce the transcriptions did not include the initial 12 minutes training time, but did include the time required for the participants to correct most of the recognition mistakes made by DNS and to add new words to the vocabulary. Although not investigated, it is expected that because DNS has a self-train feature its performance (number of words correctly recognised) would improve with use. To

defeat the self-train feature at the start of each transcription session the DNS system was reset by, initializing the DNS memory and, selecting a new user profile before each participant trained DNS. It should be noted that the times taken and the number of correct recognitions for the participants are only roughly of the same order to those obtained during the manual transcriptions described in Section 6. Further studies using larger sample sizes are needed to establish the validity of these results. It was also noted that when using DNS the amount of actual typing required was significantly reduced as most of the input was by voice, with the keyboard only being used to make some of the more difficult corrections or to input new words into the vocabulary.

7.5 Using DNS to transcribe interviews

As mentioned above, interviews between TOA Group analysts and ADF clients are often recorded, and the usual way of using DNS when making transcriptions is to listen to the recorded speech and to then repeat what is heard via a microphone into the DNS system. To improve speaker recognition, a short period of training was required before the start of the transcription process. The results, which are presented in this and the previous section, indicate that the manual and the DNS assisted approaches are comparable, and that less typing is generally required when using DNS. The results also show that using DNS does not necessarily give a reduction in the total time required to produce a transcription.

8. Factors that Influence DNS Recognition

This section describes an investigation aimed at identifying and quantifying the factors that influence the number of words recognised during transcriptions made using DNS.

8.1 Improving the efficiency of the DNS transcription process

Using DNS to assist with the transcriptions as described above is relatively inefficient as there is still a need to insert a human in the process chain to correct errors as they are encountered. To improve efficiency, what is really required is for the manual part of the process to be eliminated. That is, for the output of the recorder to be fed directly into the DNS system without manual intervention. Unfortunately, to achieve high recognition rates, it is still necessary to train DNS, but the only material available for training, if the transcription is done automatically after the interview has taken place, is the speech that was recorded during the interview. Training using tape recordings is not really satisfactory as there is no opportunity to stop the tape as it is being replayed to make any necessary corrections, as there is when the training is done using a human operator. Training from recorded material is not the way that DNS should be used if

the best possible recognition results are to be achieved (DNS 1999). However, it was unclear, at this stage of the study, just how successful this type of training was. It was also unclear how much some of the other factors such as SNR, vocabulary type, processor speed, computer memory capacity, and speech rate influenced the number of words correctly recognised.

8.2 Automatic speech recognition using DNS without training

This approach investigated the use of DNS without training. However, to use DNS the user is required to follow a standard set-up procedure that involves the training step, which cannot be easily bypassed and takes about 12 minutes. The standard training script '3001: The Final Odyssey' was selected, while the reference audio (see Appendix A) was input to the DNS system. No attempt was made to match the reference audio to the training script. All that was necessary was to reach the training step, select the training script, input the reference audio, and wait about 12 minutes before proceeding onto the next step. In this study, this is called the untrained state. The intention of this part of the study was to determine the intrinsic recognition capability of DNS and to identify some of the external factors that significantly influence recognition capability. The approach relies on the intrinsic ability of DNS to recognise words and to correctly transcribe them where there are no opportunities for manual intervention to correct any mistakes.

8.2.1 Physical configuration and set-up adjustments

The physical configuration is illustrated in Figure 8-1. The reference audio signal was initially captured using the audio tape recorder. To provide greater editing, filtering, and signal processing opportunities during replay, this stored reference audio signal was input via a sound card into a computer (#1) and stored as a digital audio (.WAV) file. The digital audio signal was then fed to DNS via the input to the sound card in the second computer.

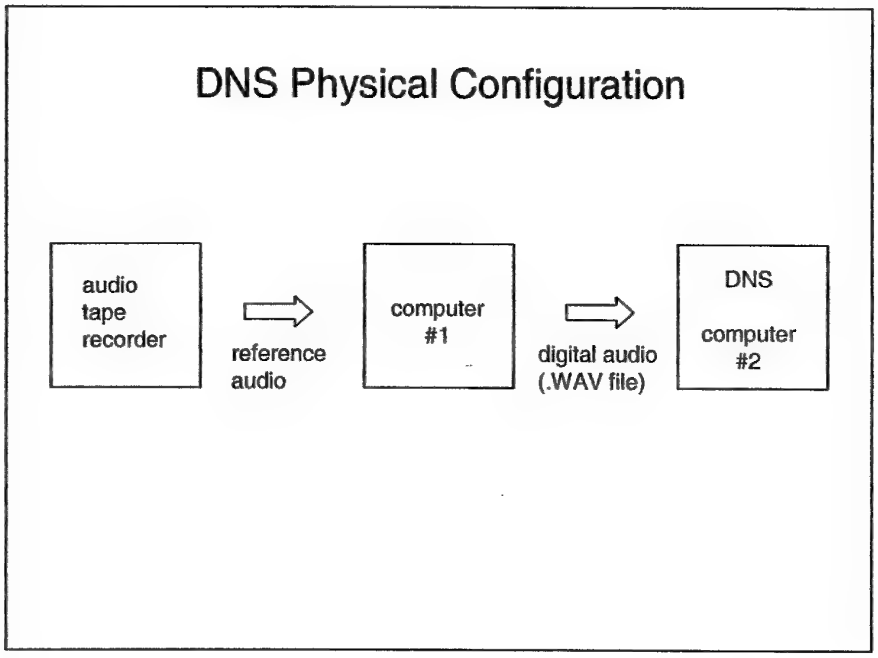

Figure 8-1 Configuration for investigating factors that influence recognition

8.2.2 DNS parametric settings and defaults

It was recognised that the parametric settings for the recording equipment, and the computer hardware and software switches, could significantly influence the results obtained. Although not all parameters were readily accessible, an attempt was made at the beginning of this part of the study to control those that were. These included those parameters associated with the tape recorder and the recorded signal, the computer software and hardware, and the DNS software.

8.2.3 Recorder and DNS input signal level

The recorded peak signal level for the reference audio, as measured by the audio level meter on the tape recorder, was 0 VU. This was the same as the input signal level that was measured at the line input to the sound card fitted to the computer (#2) that was running the DNS software. This level was well below the audio clipping level for the sound card, which is controlled by the DNS software. The line-input level to the voice recogniser is automatically adjusted during initialisation by the DNS set-up wizard, which can automatically increase or decrease the effective input level and maximise the number of words correctly recognised.

8.2.4 Recorded SNR

The SNR of the original recorded reference audio signal was 12-13 db, which is quite low, but is typical of the sound levels in an office environment in which client interviews are conducted. The background noise level, as mentioned in section 4 was 41 dbA. The SNR was measured using the DNS audio set-up wizard. Background noise, poor microphone placement, and microphone type (omnidirectional) were primarily responsible for the low SNR.

8.2.5 Computer processor speed

The recommended minimum system for DNS version 4 is a Pentium II CPU with processor speed of 300MHz and 128MB of random access memory. When installed on a computer with sufficient speed and processing capacity, DNS is capable of transcribing conversational speech, around 3 words per second (Martin and Welch 1980). To ensure that the speech algorithms could adequately accept and process the input signals in a timely manner it was necessary to install DNS on a computer that had an adequate processor speed. The CPU processor used for this part of the study was a Pentium 2 processor running at 366MHz.

8.2.6 Computer memory

The DNS active vocabulary is stored in random access memory (RAM), while the backup dictionary is stored in the hard disk. The speed of processing the speech algorithms can be increased by avoiding the use of slow memory eg, hard disk. The high-speed memory capacity of the computer needs to be sufficient to store the DNS program and any data that is generated during the transcription process. The amount of RAM used was 128Mb.

8.2.7 DNS dictionary

Recognition accuracy depends on the choice of inbuilt dictionary that is used. The only dictionary choice that was available at the start of the study was the 'UK-English' version. Later in the study, the 'Australian-English' dictionary became available and was incorporated. The result of changing the type of dictionary on the number of words recognised is described in section 9.

8.2.8 DNS default settings

Two adjustments are provided with DNS to allow the user to make a trade off between accuracy of recognition and speed of recognition. The first type of adjustment internally limits the time that DNS spends searching for a correct word match. The default setting, which provides a compromise between speed and accuracy, was used during this part of the study. The second type of adjustment allows alteration of the pause between phrases. Although natural speech is actually continuous and words

usually run into each other, there is often a pause at the end of a phrase. DNS provides facilities to adjust the average time duration of the pause between phrases to allow for better discrimination between the two prime modes of operation of DNS, dictation and command. The setting of the pause between phrases specifies the minimum time a user should wait after completing a phrase, before issuing a command. Although the dictation only mode was selected, the adjustment had the potential to affect the number of correct word recognitions. The range of the adjustment was from 100 to 1000ms. Initially, the default value of 250ms was used.

8.3 Results

The DNS system described above was used to transcribe the .WAV file, but during the set-up procedure DNS indicated that the SNR was too low. As a result, out of the 2500 words applied to the input of the untrained DNS, only 900 words appeared in the transcribed text output. Of these, only 71 were correctly recognised as being correctly spotted words in the reference audio. This corresponds to a recognition rate of 2.8%.

8.3.1 Manual scoring

Because the recognition rate was so low the NIST scoring software could not be used and the scoring had to be carried out manually. The transcription from DNS contained such a large number of incorrectly recognised words that it was impossible to determine sentence boundaries in order to insert the utterance identifiers used with the NIST automatic scoring tool. The manual scoring was not difficult, but the checking process was tedious and time consuming.

8.3.2 Factors affecting the recognition rate

During the initial set-up, DNS indicated that the SNR of the reference audio input was unacceptable (12-13 db). Generally, a low SNR influences the recognition rate in two ways. Firstly, the relatively high noise level makes it difficult for DNS to match words correctly. This means that DNS needs to extend the size of the list of words that it searches through. It searches in the active vocabulary, then in the backup dictionary. If the correct match is not found, the closest match is substituted. With a noisy signal there is a low probability of an exact word match. This causes extra searching, which takes additional time and can cause the transcribed text to lag the reference audio input by several sentences. If the lag is too great, possibly indicating that either the speech rate was excessive, or that the PC processor speed was too slow for the given SNR, then DNS leaves out words from the transcribed text in an attempt to keep up with the reference audio input signal. It was noted that near the end of transcription the production of transcribed text by DNS lagged the reference audio by so many sentences that the text at the end of the transcription was truncated prematurely. This is symptomatic of an audio buffer over-run condition.

8.3.3 Transcribed text and reference audio misalignment

Correct word recognitions could not be registered during the transcription process because the production of the transcribed text lagged the audio reference signal by several sentences. Although DNS records the reference audio temporarily during transcription and the audio can be replayed immediately after the transcription, on exit from DNS the audio recording cannot be saved. This is a limitation of DNS version 4. During replay, DNS moves a cursor through the screen text in steps that are approximately related to the current word position in the audio reference signal file. Screen capture software was used to record both the video and audio during the transcription in order to simplify the process of determining the number of correct word recognitions. However, because the cursor lagged the reference audio signal, this 'ease of use' method of scoring was discarded and the more tedious manual scoring method was used.

9. Improving Recognition

The results of the previous section were used as a basis for investigating the different factors that influence the word recognition capability of DNS for pre-recorded speech. Although DNS was the specific tool used for this study, many of the findings may be applicable to other brands of speech recogniser.

9.1 Factors that influence speech recognition using DNS

A review of the literature indicates that many factors can influence speech recognition (Lea 1982; Lea 1983; Pallett 1985). In this study only a few of these factors were investigated. The choice was based mainly on the ready availability and accessibility of inbuilt adjustments, environmental factors, computer hardware, and DNS software accessories. The factors investigated were:

- length of the pause between phrases;
- rate of speaking;
- SNR;
- impact of training;
- computer speed;
- computer memory size (RAM and hard disk); and
- dictionary type (accent).

9.2 Overview of the setup and ordering

Each of these factors was examined to determine how it influenced the number of correct word recognitions. Because there were several variables to be investigated, and the study was exploratory in nature, it was necessary to choose suitable setup

conditions and an appropriate ordering schedule in which to progress the direction of the study.

The setup of the DNS system, used the default (parametric) settings, and as far as possible only one variable was changed at a time. The recorded audio reference signal was the same as that used previously (section 8). Initially, the DNS system was untrained and used the UK-English reference vocabulary. The DNS system computer hardware was fitted with a Pentium 366 Mhz processor and 128 Mb of random access memory. The initial operating system was Windows 95. Other applications that may have required concurrent CPU processing time, or high-speed memory (RAM) access, were removed from the system.

The first set of investigations examined separately the influence of the length of the pause between phrases, the influence of the speech rate, and the influence of the SNR on the recognition rate for an untrained DNS system. The system was then trained and these factors were investigated again.

It was noted early in the study that with the untrained DNS system, and a low SNR, the production of transcribed text lagged the reference audio and that there was a large number of words lost during the transcription process. At the time, this word loss was thought to be due to the relatively slow processing speed of the computer that was used. To improve the number of words retained a faster processor, and a greater amount of memory, was used.

During the latter stages of this part of the study, the UK-English vocabulary was replaced with an Australian-English vocabulary.

9.3 Length of the pause between phrases

DNS converts speech into text, or into commands that can control the operation of the computer. Because of this, when using DNS it is important to clearly distinguish between the text translation mode and the computer command mode. The *Pause Between Phrases* feature in DNS specifies the minimum amount of time that a user should wait, after completing a phrase, before issuing a command. For this part of the study, DNS was given no training and the *Dictation Only* mode of DNS was selected, while the *Pause Between Phrases* setting was varied. Two settings used were, 100 ms and 250 ms (the default value). Again, as in the previous cases, the input signal was the reference audio. Table 9-1 shows the number of correct recognitions for pause between phrases settings of 100 ms and 250 ms. From the table it can be noted that there is little difference between the two settings. Because of this, the *Pause Between Phrases* adjustment was fixed at 250 ms (default setting) for most of the remainder of the study. Manual scoring was used to produce the results.

Table 9-1 Correct recognitions vs pause between phrases

Pause Between Phrases	Correctly Recognised Words
100ms	73 words in 2500 (2.9%)
250ms (default)	71 words in 2500 (2.8%)

9.4 Rate of speaking

CEP digital audio editing software was used to decrease the replayed speech rate (CEP has a feature that allows for the pitch independent time extension of audio files). The rationale for slowing the speech was that slower speech would give the computer programmed with the DNS software more time to process each word and phrase, and hence increase the recognition rate. The basic physical configuration for the CEP setup (computer #1) is shown in Figure 8-1.

The speech rate for the reference audio was reduced in two steps, firstly by 10% and then by 20%. The results of slowing the audio rate are shown in Table 9-2. From the table it can be seen that for the untrained DNS the 10% change produced virtually no measurable improvement in the number of correct word recognitions, while the 20% reduction produced only marginal improvements.

Table 9-2 Correct recognitions vs speech rate reduction

Audio Speech Rate Reduction	Speech rate (words per second)	Correctly Recognised Words
Reference audio	3.3	71 words in 2500 (2.8%)
Audio slowed by 10%	3.0	73 words in 2500 (2.9%)
Audio slowed by 20%	2.7	97 words in 2500 (3.9%)

9.5 SNR

The reference audio SNR (approximately 13 db) that was used (refer to section 8) was unacceptable for use with DNS (during set-up, DNS provides the user with an indication of the suitability of the SNR of the input signal). It was noted that the minimum SNR threshold required by DNS was approximately 15 db, below which the signal to noise level was unacceptable.

Digital filtering (CEP) was used to improve the SNR by filtering (removing) some of the noise. As the noise and the reference audio were combined during the recording process, the amount of noise filtering that could be usefully applied before significantly impairing the quality of the signal, was limited. However, reducing the level of noise by 3 db improved the SNR of the reference audio sufficient (16 db) to obtain an acceptable signal indication from the DNS system. This resulted in a small increase in the number of words that were correctly recognised from 71 to 118. See Table 9-3.

Table 9-3 Correct recognitions – with and without noise filtering

Reference Audio	SNR	Correctly Recognised Words
Reference Audio	13db	71 words in 2500 (2.8%)
Filtered Reference Audio	16db	118 words in 2500 (4.7%)

9.6 Altering the SNR and speech rate together

Changes to the SNR and the speech rate were then carried out simultaneously. The SNR was improved by 3db from 13db to 16db, while the speech rate for the reference audio, 3.3 w/s, was reduced by 10% and then 20%. The pause between phrases was maintained at the default value of 250 ms. Table 9-4, which for direct comparison includes the results given in Table 9-2 and Table 9-3, shows the results of reducing the speech rate while improving the SNR by 3db.

Table 9-4 Correct recognitions – untrained recogniser, low SNR audio.

Speech Rate Reduction	SNR	Correctly Recognised Words
Ref audio (3.3 w/s)	13db	71 words in 2500 (2.8%)
Ref audio slowed by 10%	13db	73 words in 2500 (2.9%)
Ref audio slowed by 20%	13db	97 words in 2500 (3.9%)
Filtered reference audio	16db	118 words in 2500 (4.7%)
Filtered reference audio slowed by 10% (2.97 w/s)	16db	46 words in 2500 (1.8%)
Filtered reference audio slowed by 20% (2.64 w/s)	16db	51 words in 2500 (2.0%)

For each of the row entries shown in Table 9-4, the recognition rate remains relatively low (less than 5%). Note that the number of correct recognitions dropped when both noise reduction and speed reductions of 10% and 20% were applied. The reasons for this are unclear, but the combination of the noise filtering and speech rate reduction could be responsible for a possible change in the acoustic characteristics of the reference audio signal.

9.7 Impact of training

9.7.1 Training process

The recognition rate of DNS can be improved with training. Speech training is the process of inputting into a speech recogniser a representative sample of audio speech that is associated with a carefully prepared matching text. The training helps DNS to interpret the individual voice nuances associated with different speakers. Instead of using the short passage of text that was supplied with the DNS application for training, the reference audio and matching reference text, used in the previous part of this study, was used. As the reference audio was recorded, there was no opportunity for human intervention to correct errors. The training on this material took 12 minutes and 30 seconds, the duration of the recording. As some of the words in the reference text were not part of the DNS vocabulary, they needed to be added to the DNS vocabulary prior to training, otherwise they would not be recognised.

9.7.2 Results after training

At the completion of the training the reference audio was then input to the DNS system. It should be noted that this was the same reference audio that was used to train DNS, and that in normal operation it is not usual to test the DNS system using the training material. With training, the number of correct recognitions increased to 339 (13%), up significantly from the untrained case.

Following this, the speech rate was reduced by 10% and then 20% and the number of recognitions increased to 759 (30%) and 1172 (47%) respectively. Each time the speech rate was reduced, DNS was retrained from scratch using the new speech rate, thus ensuring that the effects of training were not accumulative.

Using filtering to improve the SNR of the reference audio input also increased the number of recognitions to 1427 (57%), and reducing the speech rate gave further improvements. A 10% reduction in the speech rate produced 1598 (64%) recognitions. However, a further reduction in the speech rate to 20% decreased the number of recognitions to 1548 (62%). This last result indicates that that there is a limit to the extent that the speech rate can be slowed. The combined results are shown in table 9-5.

Table 9-5 Correct recognitions for DNS system trained on recorded material that was subsequently used for testing

Input	SNR	Correct Recognitions
Ref audio	13db	339 words in 2500 (13%)
Ref audio slowed by 10%	13db	759 words in 2500 (30%)
Ref audio slowed by 20%	13db	1172 words in 2500 (47%)
Reference audio filtered	16db	1427 words in 2500 (57%)
Ref Audio slowed by 10% and filtered	16db	1598 words in 2500 (64%)
Ref Audio slowed by 20% and filtered	16db	1548 words in 2500 (62%)

It was noted that for the trained DNS system the transcribed text output still lagged the reference audio and that there were missing sections in the transcribed text. However, this situation gradually improved as the recognition rate increased. At this stage it was considered that the problem of the transcribed text output lagging the reference audio input was potentially related to limitations in computer processor speed and to a low SNR. An extensive vocabulary search requires a fast processor, and a relatively low SNR can make word searching more extensive, and time consuming, before a best match is found.

9.8 Type of dictionary

The speaker whose voice was used to produce the reference audio had an Australian accent so using the UK-English vocabulary was envisaged to produce less than optimum results. However, during the study a new version of DNS was released which included an Australian-English vocabulary. Using this Australian-English vocabulary resulted in an increase in the number of correctly recognised words. The results of using the different vocabularies are shown in Table 9-6. An examination of the results shows that the greatest improvement occurred when the dictionary type was changed for speech having a relatively low SNR. The implications of this finding could be examined in future studies.

Table 9-6 Recognitions for Australian-English and UK-English vocabularies

Input	SNR	UK-English	Aust-English
Reference Audio	13db	339 (13%)	1234 (49%)
Audio slowed 10%	13db	759 (30%)	1346 (54%)
Audio slowed 20%	13db	1172 (47%)	1455 (58%)
Audio filtered	16db	1427 (57%)	1669 (66%)
Audio slowed 10%	16db	1598 (64%)	1730 (69%)
Audio slowed 20%	16db	1548 (62%)	1825 (72%)

For the remainder of the study the Australian-English dictionary was used.

9.9 Computer speed

As mentioned previously, the recommended minimum system for DNS version 4 is a Pentium II CPU with processor speed of 300MHz. In an attempt to correct the problem of the transcribed text lagging the reference audio signal, a computer fitted with a Pentium 3 processor running at 800 MHz was substituted for the initial Pentium 2 (366Mhz) processor configuration. The operating system was also changed from Windows 95 to Windows NT 4.0. As a result of the change, the total number of words in the transcribed text did increase. However, the faster processor did not noticeably increase the number of words correctly recognised. The faster processor was retained for the remainder of the study.

9.10 Computer memory

At this stage DNS memory usage during transcription was also investigated to discover its impact on word recognition rate. For the study, the computer containing Windows NT had only one application installed, DNS. The resulting memory usage was primarily a function of the memory requirements of the operating system and the DNS application. The 'Performance monitor' application in Windows NT was used to investigate memory usage during the study. The results are shown in Figure 9-1.

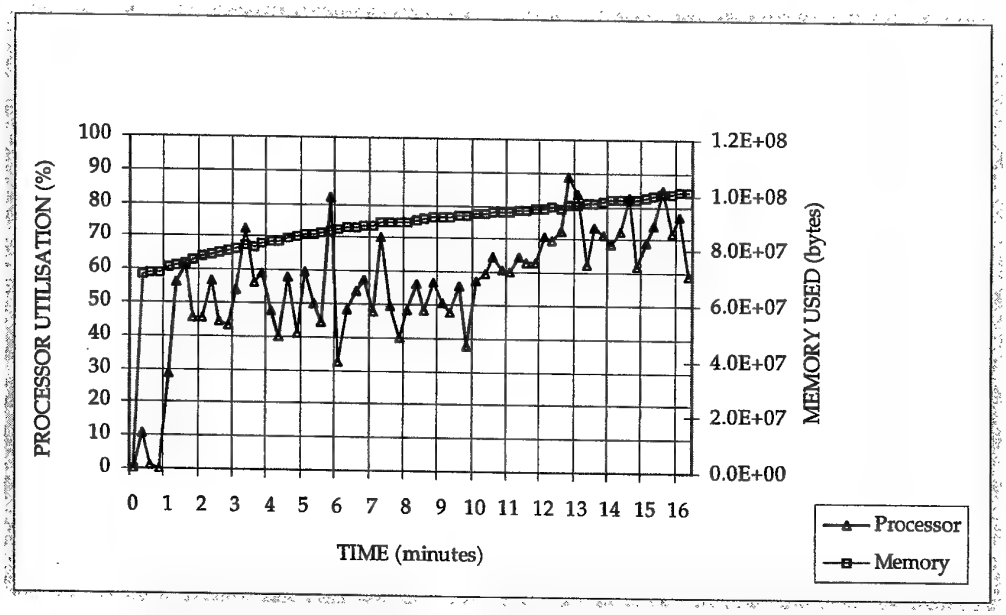

Figure 9-1 Processor utilisation and memory use by DNS

With DNS loaded in memory (RAM), but not transcribing, the memory used (committed bytes) is static at around 70 Mbytes. With DNS transcribing, additional

RAM memory is consumed at the rate of 30 Mbytes every 16 minutes. Since the audio used in this part of the experiment did not extend beyond about 16 minutes (with a 20% time extension of the digital audio files) the effect of limited memory capacity on the number of correct word recognitions could not be determined. However, it was estimated that after approximately 30 minutes of transcription the 128 Mbytes of RAM memory would have been exhausted leading to a potential degradation of the performance of the DNS system because of the need to access slow memory (hard disk). This study used a relatively short duration script with a limited number of words. The amount of memory used with longer scripts and the resulting impact on performance could be the subjects of further studies.

10. Combining the Improvements

This section describes the next stage of the study in which some of the factors that led to an increase in the number of words recognised are combined.

10.1 Overview

An examination of the results described in the previous section indicated that each of the factors reported on in the previous section had an influence on the number of words correctly recognised. However, some of those factors (SNR, speech rate, and training) appeared more significant than others.

10.1.1 SNR

The original reference audio was recorded with a SNR of 13 db, which initially was too low for DNS to produce suitable results. The post-recording audio filtering that was performed on the reference audio signal improved the SNR by 3 db to 16 db, and this significantly increased the number of correct word recognitions. However, 16 db is only marginally above the noise threshold (15 db) for DNS. An independent study by Littlefield (Littlefield and Hashemi-Sakhtsari 2002), that more closely examined the effects of noise around this noise threshold region, showed that increases in the number of words recognised were possible by increasing the SNR even further. As the original reference audio was in the form of a recording, the only practical way to significantly increase the SNR, and maintain the speech fidelity, was to produce a new low-noise reference audio recording.

10.1.2 Lowering the speech rate

Lowering the speech rate also increased the number of words recognised. However, at this stage of the study, the extent to which the speech rate could be lowered, but still provide improvements, was still unknown. In section 9 it was shown that with the aid of signal processing and filtering a reduction in speech rate by 10% provided an increase in the number of recognitions, but that reducing the speech rate by 20%

caused the number of recognitions to start to decrease. Continually decreasing the speech rate, while still maintaining the pitch, involves altering the audio signal and eventually leads to unacceptable audio distortion. The impact of the reduction of speech rate on word recognition could be the subject of a future study.

10.1.3 Training

In section 9.7 it was shown that training produced a significant increase in the number of words recognised. However, it should be noted that the training was carried out using the pre-recorded audio reference material (Appendix A). The same script containing all the words to be recognised by DNS during the testing phase had previously been used during the training phase. This is not the way in which the designers of DNS intended it to be used. Usually, the training material is different to the testing material. It should be noted that for the transcription tasks that are the focus of this study, as training requires an accurate transcript of the audio reference and the production of this transcript can be onerous, there seems to be little practical benefit to be gained by training DNS using any part of the recorded reference audio material. A later part of this study (section 11) investigates the performance of DNS in transcribing unseen material.

10.2 Setup and ordering

10.2.1 Developing a new audio reference and reference text

To increase the SNR a new audio reference recording was produced. This recording process followed that which was used previously and used the same speaker. However, a close-talking noise cancelling microphone of the type associated with a dedicated voice recognition headset was used. The original reference text was also used. The new recording and the reference text were then compared and the reference text was updated to account for the small differences that were produced during the dictation and recording process. The reference audio was then converted to a digital audio file using the CEP editing software. The resulting reference text consisted of a total of 2485 words and the duration of the recording was 13 minutes and 4 seconds, giving a corresponding speech rate of approximately 3.2 words per second. The resulting SNR was 26db, considerably higher than was previously (13db) used.

10.2.2 Training

Investigations using the higher SNR were then carried out to determine the impact of training on the number of words correctly recognised. Varying amounts of speech rate reduction were investigated with DNS in the untrained state and then in the trained state.

10.2.3 Process

For the DNS without training, the reference audio was input and the number of correctly recognised words was counted. Noise filtering was not used. The reference audio was then slowed by 5%, and then 10%, and the number of correctly recognised words was counted again. Table 10-1 shows the results. Note that when the reference audio is slowed by 10%, the number of recognitions is less than that for the 5% case. The reasons for this are unclear, but could be due to signal distortion that may occur during speed reduction.

The process was repeated with DNS trained using the newly recorded high SNR reference audio. The training took 13 minutes and 4 seconds. Noise filtering was not used while the reference audio was slowed by 5%, 10%, 15%, and 20%. The results are shown in Table 10-1. Beyond 20% the number of recognitions decreased (not shown in the table).

Table 10-1 Correct recognitions – for DNS training and slowed audio (SNR 26db)

Input	Training State	Correct Recognitions
Reference Audio	No Training	1495 words in 2485 (60%)
Audio slowed by 5%	No Training	1570 words in 2485 (63%)
Audio slowed by 10%	No Training	1446 words in 2485 (58%)
Reference Audio	Trained	2178 words in 2485 (87%)
Audio slowed by 5%	Trained	2242 words in 2485 (90%)
Audio slowed by 10%	Trained	2263 words in 2485 (91%)
Audio slowed by 15%	Trained	2290 words in 2485 (92%)
Audio slowed by 20%	Trained	2302 words in 2485 (92.6%)

10.3 Results of combining the improvements

The results show that for an untrained DNS with a high SNR (26 db) the word recognition rate for the reference audio is 60%. The production of text by DNS in this case lagged the audio by 2 or 3 sentences. Slowing the speech rate by 5% increases slightly the recognition rate (63%). In this case the text produced by DNS occurred synchronously (no lag) with the audio. A further reduction in the speech rate by 10% caused a decrease in the number of words recognised (58%).

With training, the reference audio resulted in 87% of the words being correctly recognised. Recall that training (and testing) was carried out using only the recorded audio reference. When the reference audio was slowed by 5%, 90% of the words were correctly recognised. Thereafter, when the audio was slowed in 5% steps, the number of words correctly recognised increased by approximately 1% for each step.

These results show that a high SNR and training using recorded material were significant factors in influencing the number of words correctly recognised. Reducing the speech rate had less influence on the number of correct word recognitions. Using a faster processor did increase the number of words transcribed, but did not increase the recognition rate.

11. Manual Transcription Aided by DNS

This section describes the next part of the study, which was to determine the utility of an untrained DNS in assisting manual transcription.

11.1 Aim

The previous section showed that transcriptions using an untrained DNS system and reference audio with a high SNR resulted in approximately 60% of the words being correctly recognised. However, although training improves recognition it takes time and it may not always be possible. What is of some interest to operational analysis within TOA Group is how an untrained DNS system may practically help the manual transcription process.

11.2 Process

For this part of the study an untrained (see section 8.3) DNS system was used to transcribe the new recording of the reference audio. Two people participated in this part of the study. Each participant was independently tasked to correct any errors in the transcription with the aid of DNS while listening to the replayed audio. Time and accuracy measures were used to assess the performance using this approach. Set-up time and the time required for DNS to automatically carry out the initial transcription were not included.

11.3 Results

As can be seen from table 10-1 in the previous section, the untrained DNS system was able to correctly recognise approximately 60% of the words. In this part of the study, the participants were responsible for correcting the remaining errors with the assistance of DNS. The results are shown in Table 11-1. These results (90 and 95 minutes) show that the time taken to correct the errors with the assistance of DNS is generally longer than it takes to correct errors using only the manual transcription method (see section 6). However, there were only two participants and this finding is not conclusive. Typing experience and familiarity with the DNS system are seen as factors that can significantly influence the results, and further studies are needed to establish the relative merits of each approach. Again, the word accuracy was also less

than 100%; due mainly to the small number of residual errors made by the participants. This also was essentially a one-pass error correction process.

Table 11-1 Untrained DNS assisted transcriptions

Participant	Method	Time taken (mins)	Word accuracy (%)
1	Untrained DNS	95	98
2	Untrained DNS	90	98

12. Extending the Reference Text

The results in section 10 show that the word recognition rate for an untrained DNS is about 60%, and that training DNS can help to significantly increase the recognition rate. However, producing a DNS training script from the recorded material can require a considerable amount of manual effort, particularly if all of the recorded material is used to produce the training script. Benefits are likely to arise when the training material is relatively short compared to the total length of the material to be transcribed. This section describes the process used and the results obtained from extending the length of the reference audio, but only using a portion of the reference audio to train DNS.

12.1 Setup and process

A new reference text and matching audio recording of 30 minutes was produced. The audio recording source was chosen from a random selection of broadcast transcripts that were available from the Australian Broadcasting Corporation's Radio National web site. This source was re-recorded by the same speaker as was used previously. However, the new reference audio SNR was 22 db, due to a slightly different microphone placement. The reference text was then corrected to match the 30 minutes recording. The total number of words in the new reference text was 4975.

For the untrained DNS, the standard training script '3001: The Final Odyssey' was selected, and the complete 30 minutes of reference audio was transcribed by the untrained DNS. No attempt was made to match the reference audio to the training script. The results are shown in Table 12-1.

DNS was then trained with a 12 minutes section of the reference audio, which consisted of the first 1929 words of the new reference text. The average speech rate for the 30 minutes period was approximately 2.8 w/s.

As the number of transcribed words to be checked was considerable, manual methods of counting the number of correct words would have been onerous. To assist the

counting, the NIST Sclite scoring software was used to compare the reference and transcribed text files. The use of NIST Sclite scoring software required the insertion of utterance identifiers and new-line characters at sentence boundaries in both the reference text and the transcribed text.

12.2 Results

The results are shown in Table 12-1. From the table it can be seen that, with no training and a SNR of 22 db, DNS correctly recognised 61% of the words in the 30 minutes of reference audio. With training (12 minutes), the recognition rate for the full 30 minutes of reference audio increased to 69%. For the 12 minutes section that was trained, DNS correctly recognised 79% of the words, and for the remaining 18 minutes the recognition rate dropped to 59%, just below the recognition rate for the untrained case. The implication of this finding is that training using recorded material does not increase the recognition rate on unseen material.

Table 12-1 Percentage correct recognitions with and without training

Input	30 mins audio	With 12 mins training	Remaining 18 mins
No training	61%	-	-
Trained	69%	79%	59%

13. Conclusions.

The study aimed to provide an indicative assessment of the effectiveness of speech recognition software in facilitating the process of transcribing audiotapes. The commercial speech recognition software tool investigated during the study was DNS. By design, the approach to the study was exploratory in nature and did not encompass the complete rigour associated with a strict scientific experiment. Rather, the intention of the study was to identify, and scope, some of the important issues that could be explored more fully in future studies.

Manual transcriptions of recorded material (a simulated interview), unaided by DNS, produced word recognition accuracies of up to 96%, and the time taken for these transcriptions ranged between 70 and 83 minutes. Higher accuracies require checking that is beyond that usually achieved during a one-pass transcription process. Transcriptions of recorded material using a DNS system trained to recognise the voice of the person transcribing, where the person listened to recorded speech and repeated what was heard into a microphone, produced recognition accuracies as high as 96%, with times of 74 minutes and 110 minutes. During these DNS assisted transcriptions, the errors were corrected and no additional, post transcription, error checking was carried out by the participants. Again, this was a one-pass transcription process. The

results for both types of transcriptions are similar, however, these results are largely dependent on the typing and transcription related experience of the participants.

Of interest was the intrinsic ability of DNS to transcribe recorded material and several factors that influence speech recognition rates of DNS when transcribing recorded material were investigated. A factor that was found to be most significant was the SNR. It was shown that the higher the SNR the greater the number of words recognised. For a SNR greater than 26 db excellent speech recognition rates were achieved. However, below about 16 db the number of words recognised was quite low. Other factors also improved the speech recognition rate. Using an Australian-English vocabulary, slowing down the speech, optimising the gap between phrases, and making sure that the processor speed and memory capacity were sufficient, all resulted in improvements in the speech recognition rates.

Training also improved the DNS recognition rate for transcriptions that used the same recording material during testing as that used during the training. The best recognition rate achieved with training and the above factors optimally adjusted was just over 92%. However, without training the recognition rate was 60%. It was noted that training from a recording provided no improvement in the recognition of unfamiliar material - that is, training using only recorded material appears to be of no particular benefit.

DNS has not been used directly by TOA analysts to capture speech during client interviews. To successfully use DNS to capture speech during an interview it would be necessary to spend five to ten minutes training the recogniser by having each interviewee read aloud a prepared sample script. However, as most interviews take between 30 and 60 minutes, the DNS set-up and training time would significantly add to the time allocated for each interview. In an attempt to minimise the amount of disruption to the work of busy ADF personnel, speech recognition approaches, where the training of a speech recogniser is required, have not so far been used during client interviews. Usually, the conversations are recorded and manually transcribed later. However, in the future when voice recognisers are more sophisticated and require less time to train, the voice recognition approach may become more acceptable to the clients.

14. Recommendations

This study provided some useful outcomes that can be used to guide future research into the speech technology methods that could be used to improve the interview and the transcription work practices of TOA Group. The following are recommended:

- The data and observations obtained from comparing the manual and DNS assisted transcriptions indicate that the two processes are roughly comparable.

However, no comparisons were made using typists who have significant transcription experience or experience using speech recognition systems. Another limitation of the study was the small number of participants who were drawn from people who were available within C2D. Also, an assumption was made that the typing and transcription skills and abilities were representative of office workers in the general population. To provide more investigative rigour and greater statistical confidence in the results it is recommended that for future studies much larger sample sizes from a wider population be used. Larger sample sizes would help to validate any assumptions made regarding the skills and abilities of the participants involved in future studies.

- As the results obtained were for a single speaker, they are unrepresentative of an actual interview situation where there are at least two speakers. It is recommended that future work should investigate the ability of speech recogniser systems to differentiate between two or more speakers. Specialised hardware and software items will need to be constructed to assist in differentiating the voice profiles of different speakers.

- A high recognition rate for transcriptions requires a high SNR. To assist in obtaining a high SNR during the recording of an interview it is recommended that greater use be made of close fitting directional (noise cancelling) microphones such as those associated with dedicated voice recognition headsets. Alternatively, two separate microphones and associated separate channel recorders could be used. The use of omnidirectional microphones located at a distance of approximately one metre from interview participants is strongly discouraged if transcription is to take place using voice recognition software. Although not always possible, the use of quiet environmental conditions is recommended to help to keep the background noise levels to a minimum.

- DNS is only one of a range of different types of speech recognition system that are continually being improved. During the course of this study, two new versions of DNS were released. However, in order to provide a firm basis for comparison, the version used during this study was not changed (except in the later stages to incorporate the Australian-English vocabulary). It is recommended that in any future studies, a later version of DNS be used. Although DNS represents an affordable, state of the art, speech recognition product, it would also be useful to compare its strengths and weaknesses with other speech recognisers.

15. Acknowledgements

The authors are grateful to all of the members of the C2D administrative staff who participated in the study, and in particular to Justin Fidock of TOA Group who helped to produce the reference audio, and to Oliver Carr of HSI Group who converted the NIST scoring software so that it was suitable for use in a DOS environment.

16. References

DNS (1999). Dragon NaturallySpeaking User's Guide, Dragon Systems.

Fiscus, J. (1998). Sclite User Manual, US National Institute of Standards and Technology.

Kempt, T., M. Schmidt, et al. (1999). Strategies for Automatic Segmentation of Audio Data, Interactive Systems Laboratories, ILKD, University of Karlsruhe, 76128 Karlsruhe, Germany.

Lapworth, J. and N. Jager (2000). Project Vocoder, Australian Defence Force Academy.

Lea, W. (1982). What Causes Speech Recognisers to Make Mistakes. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Paris, France.

Lea, W. (1983). "Selecting the Best Speech Recogniser for the Job." Speech Technology 1(4): 10 -29.

Littlefield, J. and A. Hashemi-Sakhtsari (2002). Effects of Background Noise on an Automatic Speech Recogniser, Defence Science and Technology Organisation.

Martin, T. B. and J. R. Welch (1980). Practical Speech Recognisers and Some Performance Effectiveness Parameters (Part 1 - Motivation and General Review). NJ, U.S.A., Prentice-Hall.

Pallett, D. (1985). "Performance Assessment of Automatic Speech Recognisers." Journal of Research of the National Bureau of Standards (USA) 90(5): 1-17.

Talintyre, J. (1996). "The Listening Phone." IEE Review: 151-154.

Vozzo, A. (2001). Experimental Observations from the use of Speech Recognition Software in a Military Environment. Defence Human Factors Special Interest Group, DSTO - Edinburgh.

Appendix A

Sample text - modified transcript from an actual interview

When we actually set up our HQ now it takes a little bit longer. To get the radios sorted out, probably adds an extra half an hour, but it is usually a lower priority activity given the tactical scenario.

How long are you usually there for?
It varies depending on the tactical situation. We can be in one location for 10 minutes or ten days, we don't know, it depends.

Could we establish what we are going to compare with, what your basic system is?
Manual system. I have only used it twice so far, once was during a CPX conducted here in the barracks earlier in the year and the other time was briefly during an FTX and I found it effective.

Are you still using battle maps at the briefings or are you actually using the 'New System' to get all your LOCSTATs?
We are still using battle maps, it is probably easier during a briefing to do a battle map reference, because you are usually talking to about 20 people at the same time. Whereas referring to a screen is not that easy. So far as the detailed planning goes within the S3 cell, the plans cell, we could probably rely more in the future on the 'New System' because we'll get better at it.

Do you have much to do with overlays?
Had a little bit to do with them, mainly in the supervisory role, but the actual operator application is done by subordinates. I haven't personally done the 'New System' course yet.

Have you been invited to do the course?
Yes, I was invited but just too busy to attend.

If it was conducted away from the Base, do you think that would be easier?
No it is probably easier if it is conducted here, but when it was conducted here it was a very busy time for all of the S3 cells in the brigade, trying to get exercises planned. We just didn't get the time to go, got most of my subordinates to do it but I didn't make it myself.

When is a good time to train?
For the 'New System' training, the first two months of the year. The end is not usually worthwhile because you have people preparing to go away on leave and towards the end of the year you will find that a lot of people will be posted out of the unit that would undertake the training. Beginning of the year when you have all the new people

come in to the unit into the new positions is the optimum time. I still maintain a paper log based on radio voice messages that come in. It is not too bad, one of the points I would probably make is that we need a hand over capability built into it.

How does the set up time affect mobility?
It is dependent on the type of set up. Whether for a non-tactical situation or for a tactical situation. It depends on the notice to move, I mean if we are on a short notice to move, we don't have the time to set it up, because you are worried about too many other things at once. From what I have seen of it, it is a great concept, but I haven't used it that much because we have only had two exercises. One was a CPX where we didn't physically have people moving and the other one was a very short stint in which there wasn't manoeuvre occurring at the time. So I can't really say too much on that.. The Battle maps that I have seen so far, the detail has been great, the problem is that it is a one-person operation at the moment, because only one person can look at the screen at once. For briefings you can, you have still got to go back to paper. You have still got, you have still got to go back to big maps, because a briefing is about 20 people all looking at it.

That is the thing, isn't it, yours is outside, not as in a HQ inside a briefing tent or something like that?
For a brigade HQ, probably a great thing, where they are static and they have the facilities, they have the power generation, they have got the time and the people to set it up. For a mobile Regimental HQ, moving tactically, you don't have time to do that, you throw a map out and go.

Do you receive your overlays from Brigade HQ, is that all done by SDS?
SDS, usually, we'll receive FRAGOs with just LOCSTATs on them usually, but so far as my experience is we get an initial overlay for the activity, maybe an overlay for any FRAGOs, but other than that, not much else. For plans, we usually have to transpose it across to a map, to indicate the CO's intent or scheme of manoeuvre, it has to go to a map, because it is just too difficult at the moment to get across. When we start getting into the situation of dispersed HQ, where we have a squadron HQ miles away from us, if they are within the 'New System' net, then that is going to be a good thing. The CO can pass his intent that way. No matter which level you go to it has got to get to a point at some stage where it has got to go back to a map.

Do you think it is a useful tool for the Commander if he is out and about? If he is out, he can walk into the Squadron Commander and have a look and he can get an over all idea of where everybody is.
In that respect, it is a great tool, but the point to be made is that he is mobile as well. Unless he is receiving information on the move, then by the time he sets up and gets the information it has probably passed the time where he would receive it by radio. His reliance on the radios is easier because it is easier for him to speak to the squadron commanders and say 'where are you, what are your LOCSTATs?' versus stopping, setting up his 'New System' and waiting for the feeds before finding where they are.

So if you have got Comms on the move, if you have got the data coming across, then it is fine for that sort of thing?
That is right, if it is updating as it is going, then that will be a good thing. When he has got it to move, stop, lift it up and grab it all again, it is just as quick to ring up my HQ, which I run for him, and he says 'where are they?'

What about as far as the Command Admin nets are concerned, could you see yourselves sticking with voice on the Command and for contact reports and all those sort of things and perhaps some orders. Or would you want to move some of the orders across to say a data net and put all your admin through a data net?
I think you could have a mix, some things will never be replaced by voice and no one should ever try to change that.

Besides the contact reports, where else would you go, or where else would you, I should say?
Contact reports, quick attack warning orders on the move, change in plans, FRAGOs on the move, SITREPs on the move. Things like that which are developed as you are seeing them, and you don't have time to program them into a computer, but you get them across as quickly as possible. 'There is a machine gun nest on the right hand side, look out. Swing your axis of assault right'. You don't have time to stop and send that in, whereas the contact report for that, at a later point, could be sent by the 'New System'. Where I see the 'New System' developing as an improvement is in the ADMIN side of things. Where ADMIN traffic which is non-essential, it has got lead times and set times, but it is not as important as the Command aspect. Admin traffic, LOCSTATs for admin, manoeuvre, perfect, absolutely perfect for that because most of the admin organisations are halted on the ground, so they are set up and they send out their little branches to do the jobs. If they have got visibility all the time of where everybody is, then they can do their administrative logistics planning much more efficiently.

So you do see it as being quite effective there?
Yes, but as I was saying, I don't think it will ever replace voice. Those things are done statically. It is perfect for that role. I personally think the 'New System' is great, I really do like it. It has got a long way to go yet before it becomes that utopian working environment that we were looking for. A couple of things that I will mention now, if you are interested. The passage of information and the systems by which we pass data need to be looked at. There are no set forms for a lot of the reports and returns that we are using. And there needs to be a facility, I am trying to use the right words, I am not sure what they are, something that can be sent easily, that has got a lot of detail like tabulated data. At the moment the system we are using for tabulated data, I don't think it is very effective, it clogs up the net for too long. The reports and returns need to be reviewed so that we can cover the whole gamut of the Brigade.

What they are doing now is inputting all the Brigade reports from the Aide Memoir into the 'New System'. So your SITREPS, all those sort of things, are going to be in that aide memoir format that you have now in the Brigade, so is that adequate?
I would have to look at it when it comes out, but I would say yes.

Then it will be just a matter of filling out the particular fields. Because it is in a field, it is a small text file, so bang it just goes.
The other thing is the fields need to be able to be modified, for example, some of the reports and returns the Brigade requires have many serials. This unit only reports on a small proportion of them, so instead of sending a document that has all the fields in it, we send one with only the relevant fields, it reduces the time. We don't have tanks for example, so we don't send anything on tanks, but because it is part of that field, it always goes as part of the return. So the operator needs to be able to modify the fields as applicable to the operation of his unit. It is not enough just to say, we are going to put all the reports and returns on, we need to be able to modify them so they go across quicker. In the field I can demonstrate that to you quite easily, it is a bit harder to talk, a bit harder talking about it.

Yes, I know what you are talking about. That is one thing that we need to look at. So as the system stands now, how would you rate that?
I would say it is the same as what we have got at the moment because it only addresses certain reports and returns. In that respect it is the same, but the potential for it is much greater, it just needs to be worked a bit more that is all. It all comes back to tactical situation, that is a question the Brigade would be able to answer for you, because they are static most of the time. For us it depends on the tactical situation, if we have got the time, we can look at it, if we haven't got the time, if the regiment is moving, then it is no more of a help than anything else.

If the regiment is moving along and the LOCSTATs are coming in, verbally, someone is plotting their location I take it in the back of the ACV.
Yes they are, but there is a time delay between when we stop and they put it on the 'New System' and send it out. And that time is dependent on the tactical situation. Is it night time, is it blacked out in the back of the vehicle, is it low noise, therefore no movement in the back of the vehicle type situation. What are the threat levels of certain things at certain times that would preclude anything happening. Plus the fact that you are talking manpower usage, we have X amount of manpower in a HQ. When we stop there are a million things we have to do all at once to get it ready to continue with operations, we start taking people out to start plugging information in, we need more manpower on the outside for cam nets. Start to do security patrols and all that sort of thing, but that is not such a big problem.

The data is automated and coming into the system, I mean it is going straight on to the computer.
That is right, but the information out, from us, won't occur until everything else is done. Because someone has got to put it in the computer and send it. So, I have tried working with the system, just doing reports and returns on set formats that I have got

there like a warning order is in a set format. Actually typing on the move and stuff like that, I don't have a problem with that, but a lot of people can't do that either, motion sickness and things like that in the back of the vehicle. Being thrown all over the place, so I don't have a problem with it personally, but a lot of other people do.

How do you find actually entering that when you are on the move?, You said you don't have a problem with it, but with the keyboard and screen, is it something you get used to?
You get used to it, I don't have a problem with it, but I know other people that I work with suffer from motion sickness and they can't physically work in the vehicle on the computer whilst its moving, otherwise they are chucking up in ten minutes. That is just something that you should be taking into consideration, mobile doesn't necessarily mean that you can keep working, whereas you can sit on a radio and have your head up.

DISTRIBUTION LIST

Using Speech Technology to Improve Transcriptions: An Exploratory Study

Ashley Cook, Alex Yates and Ahmad Hashemi-Sakhtsari

AUSTRALIA

DEFENCE ORGANISATION

	No. of copies
Task Sponsor CC2D	1

S&T Program

Chief Defence Scientist FAS Science Policy AS Science Corporate Management Director General Science Policy Development	shared copy
Counsellor Defence Science, London	Doc Data Sheet
Counsellor Defence Science, Washington	Doc Data Sheet
Scientific Adviser to MRDC, Thailand	Doc Data Sheet
Scientific Adviser Joint	1
Navy Scientific Adviser	Doc Data Sheet & Distribution list
Scientific Adviser - Army	Doc Data Sheet & Distribution list
Air Force Scientific Adviser	Doc Data Sheet & Distribution list
Scientific Adviser to the DMO M&A	1
Scientific Adviser to the DMO ELL	1
Director of Trials	1

Information Sciences Laboratory

Chief Command & Control Division	Doc Data Sheet
Research Leader Command & Intelligence Environments Branch	1
Research Leader Military Information Enterprise Branch	1
Research Leader Theatre Command Analysis Branch	Doc Data Sheet
Head Virtual Enterprises	Doc Data Sheet
Head Systems Simulation and Assessment	Doc Data Sheet
Head Theatre Operations Analysis	1
Head Intelligence Analysis	Doc Data Sheet
Head Human Systems Integration	Doc Data Sheet
Head C2 Australian Theatre	Doc Data Sheet
Head HQ Systems Experimentation	Doc Data Sheet
Head Information Systems	Doc Data Sheet
Head Information Exploitation	Doc Data Sheet
Author (s) Mr Alex Yates, TOA Group C2D	10
Mr Ashley Cook, TOA Group C2D	1
Dr Ahmad Hashemi-Sakhtsari, HSI Group C2D	2
Publications and Publicity Officer, C2D/EOC2D	1 shared copy

DSTO Library and Archives

Library Edinburgh	1 copy & Doc Data Sheet
Australian Archives	1

Capability Systems Division

Director General Maritime Development	Doc Data Sheet
Director General Information Capability Development	Doc Data Sheet

Office of the Chief Information Officer

Deputy CIO	Doc Data Sheet
Director General Information Policy and Plans	Doc Data Sheet
AS Information Structures and Futures	Doc Data Sheet
AS Information Architecture and Management	Doc Data Sheet
Director General Australian Defence Simulation Office	Doc Data Sheet

Strategy Group

Director General Military Strategy	Doc Data Sheet
Director General Preparedness	Doc Data Sheet

HQAST

SO (Science) (ASJIC)	Doc Data Sheet

Navy

Director General Navy Capability, Performance and Plans, Navy Headquarters	Doc Data Sheet
Director General Navy Strategic Policy and Futures, Navy Headquarters	Doc Data Sheet

Air Force

SO (Science) - Headquarters Air Combat Group, RAAF Base, Williamtown NSW 2314	Doc Data Sht & Exec Summ

Army

ABCA National Standardisation Officer, Land Warfare Development Sector, Puckapunyal	e-mailed Doc Data Sheet
SO (Science), Deployable Joint Force Headquarters (DJFHQ) (L), Enoggera QLD	Doc Data Sheet
SO (Science) - Land Headquarters (LHQ), Victoria Barracks NSW	Doc Data & Exec Summ

Intelligence Program

DGSTA Defence Intelligence Organisation	1
Manager, Information Centre, Defence Intelligence Organisation	1 (PDF version)
Assistant Secretary Corporate, Defence Imagery and Geospatial Organisation	Doc Data Sheet

Defence Materiel Organisation	
Head Airborne Surveillance and Control	Doc Data Sheet
Head Aerospace Systems Division	Doc Data Sheet
Head Electronic Systems Division	Doc Data Sheet
Head Maritime Systems Division	Doc Data Sheet
Head Land Systems Division	Doc Data Sheet
Head Industry Division	Doc Data Sheet
Chief Joint Logistics Command	Doc Data Sheet
Management Information Systems Division	Doc Data Sheet
Head Materiel Finance	Doc Data Sheet
Defence Libraries	
Library Manager, DLS-Canberra	Doc Data Sheet
Library Manager, DLS - Sydney West	Doc Data Sheet
OTHER ORGANISATIONS	
National Library of Australia	1
NASA (Canberra)	1
UNIVERSITIES AND COLLEGES	
Australian Defence Force Academy	
Library	1
Head of Aerospace and Mechanical Engineering	1
Hargrave Library, Monash University	Doc Data Sheet
Librarian, Flinders University	1

OUTSIDE AUSTRALIA

INTERNATIONAL DEFENCE INFORMATION CENTRES	
US Defense Technical Information Center	2
UK Defence Research Information Centre	2
Canada Defence Scientific Information Service	e-mail link to pdf
NZ Defence Information Centre	1
ABSTRACTING AND INFORMATION ORGANISATIONS	
Library, Chemical Abstracts Reference Service	1
Engineering Societies Library, US	1
Materials Information, Cambridge Scientific Abstracts, US	1
Documents Librarian, The Center for Research Libraries, US	1
SPARES	5
Total number of copies:	**45**

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

Using Speech Technology to Improve Transcriptions: An Exploratory Study

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

Alex Yates, Ashley Cook and Ahmad Hashemi-Sakhtsari

5. CORPORATE AUTHOR

Information Sciences Laboratory
PO Box 1500
Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-GD-0399	AR-013-068	General Document	March 2004

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
E9505/23/104	LRR 01/348	CC2D	60	10

13. URL on the World Wide Web

http://www.dsto.defence.gov.au/corporate/reports/DSTO-GD-0399.pdf

14. RELEASE AUTHORITY

Chief, Command and Control Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CASUAL ANNOUNCEMENT Yes

18. DEFTEST DESCRIPTORS

Voice data processing
Speech recognition
Data processing
Data transmission

19. ABSTRACT

Modern speech technology is finding many new application areas within Defence. Transcription is one area where speech recognition technology is starting to replace manual methods. This adoption of speech technology is motivated by its potential to save transcribers significant amounts of time and physical effort. As part of their evaluation of military organisations, analysts from the Theatre Operational Analysis (TOA) Group within the Command and Control Division (C2D) transcribe recorded information that has been captured during interviews with Defence personnel. An exploratory study was conducted within TOA Group to look at ways of using commercial speech technology to assist in the transcribing of recorded interview material. This report describes the method used and the results obtained from that study. The report also compares traditional manual transcription approaches with newer approaches that make use of speech recognition technology. The qualitative and quantitative results obtained from this study will help to benchmark the utility of current commercial speech technology used for transcription.